DOCUMENT RESUME

ED 469 179 TM 034 484

AUTHOR	Roussos, Louis A.; Schnipke, Deborah L.; Pashley, Peter J.
TITLE	A Formulation of the Mantel-Haenszel Differential Item Functioning Parameter with Practical Implications. Statistical Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-R-96-03
PUB DATE	2000-09-00
NOTE	20p.; For a related document on the Mantel-Haenszel Differential Item Functioning Parameter, see TM 034 485.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*College Entrance Examinations; Difficulty Level; Estimation (Mathematics); *Item Bias; Item Response Theory; Law Schools; Test Items
IDENTIFIERS	*Law School Admission Test; *Mantel Haenszel Procedure; Three Parameter Model; Two Parameter Model

ABSTRACT

The Mantel-Haenszel (MH) differential item functioning (DIF) parameter for uniform DIF is well defined when item responses follow the two-parameter-logistic (2PPL) item response function (IRF), but not when they follow the three-parameter-logistic (3PL) IRF, the model typically used with multiple choice items. This research report presents a general formulation of the MH DIF population parameter for any IRF and presents results for numerous 3PL uniform DIF conditions. The results indicate that for items of medium or high difficulty, the 2PL DIF parameter formulation can overestimate the 3PL DIF parameter and the MH DIF estimator may exhibit less than expected power to identify even substantial DIF in certain circumstances. Implications of this study on the routine operational task of identifying DIF at the Law School Admission Council are still not known, and may in fact be minimal. However, because some items on the Law School Admission Test (LSAT) are known to exhibit guessing behavior, the results suggest that additional research is warranted. (Contains 2 tables, 5 figures, and 13 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 469 179

TM

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. VASELECK

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A Formulation of the Mantel-Haenszel Differential Item Functioning Parameter With Practical Implications

Louis A. Roussos, Deborah L. Schnipke, and Peter J. Pashley
Law School Admission Council

Law School Admission Council
Statistical Report 96-03
September 2000

TM034484

LAW Services

A Publication of the Law School Admission Council

BEST COPY AVAILABLE

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 197 law schools in the United States and Canada.

Copyright© 2000 by Law School Admission Council, Inc.

All rights reserved. This report may not be reproduced or transmitted, in whole or in part, by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, Box 40, 661 Penn Street, Newtown, PA 18940-0040

LSAT® and the Law Services logo are registered marks of the Law School Admission Council, Inc.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

When examinees from two different subgroups have the same ability distribution (or are "matched" on ability) but are not equally likely to answer a particular item correctly, the item is said to exhibit DIF (*d*ifferential *i*tem *f*unctioning; that is, the item functions differently in the two groups). When test data are analyzed, a statistical measure of DIF is calculated for each item so that items with large values of DIF (i.e., items with a large difference in the probability of equal ability examinees in the two groups answering correctly) can be investigated to determine if the item should be removed from the test and/or item pool (the group of items from which new tests are assembled). The Mantel-Haenszel (MH) procedure, which is used at the Law School Admission Council (LSAC), has become the most widely used procedure for measuring DIF and is recognized as the testing industry standard. The behavior of the MH DIF parameter is well understood for items on which no guessing occurs, but not for items where guessing does occur; often the case with multiple-choice items.

This research report presents a general formulation of the MH DIF parameter that is equally appropriate for items on which guessing occurs and for items on which no guessing occurs. The value for this parameter is calculated for numerous realistic conditions to explore its behavior in situations where DIF might occur with real data. Practitioners have assumed that the MH DIF parameter behaves similarly regardless of guessing behavior, but our results indicate that guessing can affect the parameter's value for relatively difficult items. As a result, the MH DIF statistic should be used with caution until the apparent deficiencies of this procedure are better understood or corrected.

Before items are tested empirically for DIF at LSAC, and even before they are pretested (administered to examinees for the first time), they are subjected to rigorous sensitivity reviews. Additionally, real data do not mimic simulated data exactly. Thus, the implications of this study on the routine operational task of identifying DIF at LSAC are still unknown, and may in fact be minimal. However, because some items on the Law School Admission Test (LSAT) are known to exhibit guessing behavior, the results certainly suggest that additional research is warranted.

Abstract

The Mantel-Haenszel (MH) differential item functioning (DIF) parameter for uniform DIF is well defined when item responses follow the two-parameter-logistic (2PL) item response function (IRF), but not when they follow the three-parameter-logistic (3PL) IRF, the model typically used with multiple-choice items. This research report presents a general formulation of the MH DIF population parameter for any IRF and presents results for numerous 3PL uniform DIF conditions. The results indicate that for items of medium or high difficulty, the 2PL DIF parameter formulation can overestimate the 3PL DIF parameter and the MH DIF estimator may exhibit less than expected power to identify even substantial DIF in certain situations.

Introduction

Differential item functioning (DIF) is said to occur in an item when examinees of equal proficiency (on the construct measured by a test), but from separate populations, differ in their probability of answering the item correctly. Although a large number of statistical procedures have been developed to detect DIF in test data, the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), used at the Law School Admission Council (LSAC), has become the most widely used methodology and is recognized as the testing industry standard.

The behavior of the MH DIF estimator, $\hat{\Delta}$, with respect to a number of factors, has been studied extensively in simulation studies (see, for example, Allen & Donoghue, 1996; Donoghue, Holland, & Thayer, 1993; Roussos & Stout, 1996; Shealy & Stout, 1993; and Uttaro & Millsap, 1994). However, it has not been determined how well $\hat{\Delta}$ estimates its corresponding population parameter, Δ, when a DIF item is modeled by a three-parameter-logistic (3PL) function (Birnbaum, 1968). Although the behavior of Δ is well understood when responses follow either a one- or two-parameter logistic funtion (1PL or 2PL, e.g., Donoghue et al., 1993), no general formulation of Δ has been derived, although possible general formulas have been defined (e.g., Spray & Miller, 1992). This lack of knowledge about Δ in the case of 3PL items has limited the evaluation of the statistical bias in $\hat{\Delta}$ and has also hindered the understanding of the observed effects of simulation study factors on $\hat{\Delta}$. In particular, Allen and Donoghue (1996); Donoghue, Holland, and Thayer (1993), and Uttaro and Millsap (1994) all reported that the difficulty level of a 3PL DIF item can have a sizable effect on the magnitude of $\hat{\Delta}$, but none of these studies could adequately explain the cause of this effect. Type I error studies by Allen and Donoghue (1996) and Roussos and Stout (1996) have indicated that a statistical bias is sometimes present in $\hat{\Delta}$, and that this bias varies with the difficulty level of the 3PL item being tested for DIF. (Statistical bias can be estimated in Type I error studies because the true amount of DIF is known to be zero.) Type I error bias alone, however, does not fully explain the observed relationship between $\hat{\Delta}$ and difficulty level in simulated 3PL DIF items. The purpose of this research report is to present a formulation of the population DIF parameter for the MH DIF estimator that is appropriate for any IRF model, including the 3PL model, and to describe through a systematic set of calculations the behavior of this DIF parameter with respect to a number of examinee and item factors. In particular, it will be shown that the unexplained behavior of $\hat{\Delta}$ with respect to difficulty level observed in past simulation studies can be explained, at least in part, by the behavior of the MH DIF population parameter. Moreover, it will be shown that this behavior of Δ has important practical implications for the detection of DIF in real data analyses.

Item Response Theory Terminology

Item response theory (IRT) describes the relationship between the ability or proficiency, θ, of examinees on a construct and their probability, $P_i(\theta)$, of a correct response on an item i that measures that construct. In this research report, the following notations will be used:

i = item number,
j = examinee number,
n = number of items on test,
N = number of examinees,
X_{ij} = random variable for the response of examinee j to item i (X_{ij} =1 indicates a correct response and X_{ij} = 0 indicates an incorrect response), and
$P(X_{ij}=1|\theta_j)=P(\theta_j)=$ probability of a correct response on item i for an examinee j having ability θ_j. The functional representation used for $P(\theta_j)$ is called the item response function (IRF).

Multiple-choice tests typically give an examinee four or five options from which to choose the correct response to each item. For such items, empirical observation has shown that as examinee ability decreases, the probability of a correct response does not decrease asymptotically to zero, but rather to a fairly substantial finite value, usually between 0.10 and 0.25. It is generally believed that this non-zero lower asymptote of the IRF for a multiple-choice item is due in part to examinees having a finite probability of guessing the correct answer to a

multiple-choice item due to the item format. Because of this belief, the lower asymptote parameter of an IRF is commonly referred to as the *guessing* (or *pseudo-guessing*) *parameter*. The most common parametric IRF that is used to model and simulate examinee responses to multiple-choice items is the 3PL model of Birnbaum (1968), which is given by

$$P(X_{ij}=1|\theta_j)=c_i+\frac{1-c_i}{1+e^{-1.7a_i(\theta_j-b_i)}}, \quad (1)$$

where

a_i = the discrimination parameter for item i,
b_i = the difficulty level of item i, and
c_i = the lower asymptote for item i.

When $c_i = 0$, the IRF is referred to as the two-parameter logistic (2PL) IRF. When $c_i = 0$ and $a_i = 1$ (or is constant across items), the IRF is referred to as the one-parameter logistic (1PL) IRF. The 1PL and 2PL models are often inadequate for modeling responses to multiple-choice items, thus the 3PL model is commonly used.

DIF Terminology

As stated above, DIF is defined as occurring in an item when examinees of equal proficiency, but from separate populations, differ in their probability of answering the item correctly. The item that is being tested for DIF is commonly referred to as the *studied item*. The populations of interest for DIF analyses at LSAC are based on ethnicity, gender, and geography (United States and Canada). The populations are categorized into a *reference group* population (Caucasians, males, or U.S. citizens) and a set of *focal group* populations (various minority groups, females, or Canadians). For didactic purposes we will limit our discussion to tests that are intended to measure a unidimensional construct; the situation under which the MH statistic is intended to be used. The proficiency of an examinee on the unidimensional construct will be referred to as θ.

Using the above DIF terminology, a studied item is said to display DIF when reference group examinees and focal group examinees matched on θ do not have the same probability of a correct response on the item. The most common procedure used for modeling and simulating DIF in an item is to use a different 3PL IRF for the reference group than for the focal group. The reference group IRF is denoted by $P_R(\theta)$ and the focal group IRF by $P_F(\theta)$. When the only difference between the reference and focal group IRFs is in the difficulty parameter (b_i) of the studied item, the resulting DIF is referred to as *uniform DIF* because such DIF is graphically represented as a uniform horizontal shift in the IRF for one group relative to the other (see Figure 1). All other forms of DIF are referred to as nonuniform DIF.

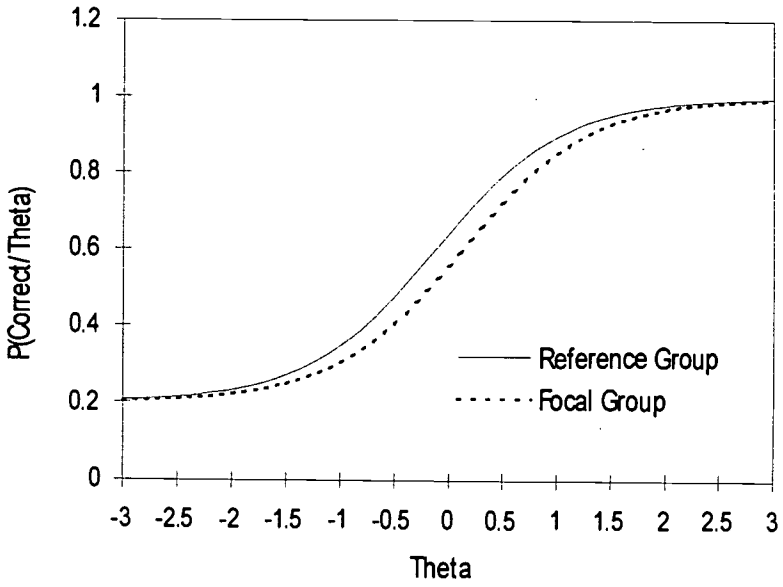

FIGURE 1. *Three-parameter-logistic item response functions for an item that exhibits uniform DIF against the focal group. For both groups, a = 1 and c = .2. In the reference group, b = –.125, and in the focal group, b = +.125.*

Let DIF(θ) be defined as the magnitude of DIF in a studied item at a particular value of θ. A variety of formulations of DIF(θ) exist in terms of $P_R(\theta)$ and $P_F(\theta)$. One common formulation parametrizes DIF(θ) as the ratio of the odds of a correct response on the studied item for the reference group, $P_R(\theta)/Q_R(\theta)$, to the odds of a correct response in the focal group, $P_F(\theta)/Q_F(\theta)$. If there is no DIF in the studied item [i.e., if $P_R(\theta) = P_F(\theta)$] then the odds ratio, denoted by α(θ), is equal to 1. In the case of uniform DIF within the framework of IRT, when the studied item is 1PL or 2PL, α(θ) can be shown through simple algebraic manipulation to be a constant across θ given by

$$\alpha(\theta) = \frac{P_R(\theta)/Q_R(\theta)}{P_F(\theta)/Q_F(\theta)} = \frac{P_R(\theta)Q_F(\theta)}{P_F(\theta)Q_R(\theta)} = e^{-1.7a(b_R - b_F)}, \qquad (2)$$

where $Q = 1 - P$ and b_R and b_F are the difficulty parameters of the studied item for the reference and focal groups, respectively. Thus, $ln[\alpha(\theta)]$ is equal to $-1.7a(b_R - b_F)$ when items follow the 1PL or 2PL model. Hence, in the case of 1PL, the log-odds-ratio at any θ is simply –1.7 times the difference in difficulty parameters for the two groups. In the case of 3PL uniform DIF, the odds ratio is not a constant across θ and is given by

$$\alpha(\theta) = \left[\frac{1 + ce^{-1.7a(\theta - b_R)}}{1 + ce^{-1.7a(\theta - b_F)}}\right] e^{-1.7a(b_R - b_F)}. \qquad (3)$$

The behavior of the odds ratio, α(θ), is important because, as will be discussed in more detail below, the MH DIF statistic is based on the estimation of similar odds ratios. Thus, it will be helpful to refer back to the previous equations when we review the MH DIF statistic and when we derive the equation for the MH DIF parameter.

The Mantel-Haenszel DIF Statistic

To evaluate whether a studied item displays DIF, examinees are first separated into reference and focal groups (for example, males and females). Next the reference and focal groups are matched on ability, ideally θ. Because θ is an unobservable variable, DIF statistics must approximate matching on θ with a matching based on the observable data at hand. The MH DIF statistic matches examinees on the basis of total test score, including the score on the studied item. Holland and Thayer (1988) have shown that when all items on a test follow the 1PL model (in which case the total test score is a sufficient statistic for θ) matching on total test score is the best approximation to matching on θ. When all the items on the test follow the 2PL or 3PL models, matching on total test score will be asymptotically equivalent to matching on θ (Stout, 1990), which suggests that for sufficiently long tests matching on total score will well approximate matching on θ. Simulation studies (such as Allen & Donoghue, 1996; Donoghue, Holland, & Thayer, 1993; Roussos & Stout, 1996; Shealy & Stout, 1993; and Uttaro & Millsap, 1994) have demonstrated the efficacy of matching on total score for 3PL items, although significant breakdowns can sometimes occur when the number of items on the test is small (for example, 25 or fewer items) and the reference and focal groups display large mean score differences (for example, one standard deviation).

After reference and focal group examinees are matched on total test score, a 2 x 2 x S contingency table is formed, where S is the number of different values of total test score. At each score level s the data can be arranged as a 2 x 2 table, as shown in Table 1.

TABLE 1
2 x 2 contingency table

	Correct	*Incorrect*	*Total*
Reference group (R)	C_{Rs}	I_{Rs}	N_{Rs}
Focal group (F)	C_{Fs}	I_{Fs}	N_{Fs}
Total group	$C_{Total,s}$	$I_{Total,s}$	$N_{Total,s}$

C_{Rs} indicates the number of reference group examinees at score level s who answered the studied item correctly. The other variables in Table 1 are analogously defined. If the item does not display DIF, the observed odds of a correct response for the two groups in each 2 x 2 table should be approximately the same because examinees in the two groups are roughly matched on ability. If the two groups do not have approximately the same odds of a correct response, the item is said to be functioning differently in the two groups (i.e., displays DIF). Thus the ratio of the reference group odds of a correct response on the studied item to the focal group odds is one natural score-level DIF estimator that can be formed from the contingency table. This odds-ratio estimator for score level s, $\hat{\alpha}_s$, is given by

$$\hat{\alpha}_S = \frac{C_{Rs} I_{Fs}}{C_{Fs} I_{Rs}}. \tag{4}$$

In the case of no DIF, $\hat{\alpha}_s$ should be approximately one for all s because the true odds for the two groups, when matched on ability, will be the same.

When $\hat{\alpha}_s$ is assumed to be estimating a constant across s, an average value of $\hat{\alpha}_s$ can be used as an overall measure of the DIF in an item. The MH odds ratio (Mantel & Haenszel, 1959), $\hat{\alpha}$, is a weighted average of $\hat{\alpha}_s$ and is given by

$$\hat{\alpha} = \frac{\sum_s \left(\frac{I_{Rs} C_{Fs}}{N_{Total,s}} \right) \hat{\alpha}_s}{\sum_s \left(\frac{I_{Rs} C_{Fs}}{N_{Total,s}} \right)}. \quad (5)$$

The MH odds ratio provides a weighted average[1] for giving stable estimation of α when $\hat{\alpha}_s$ is estimating a constant odds ratio across all score levels.

Holland and Thayer (1988) defined the MH DIF estimator ($\hat{\Delta}$) as

$$\hat{\Delta} = -2.35 \, ln(\hat{\alpha}). \quad (6)$$

This transformation of $\hat{\alpha}$ places $\hat{\Delta}$ on the Educational Testing Services (ETS) "delta scale" in the case of 1PL. The delta scale is an inverse normal transformation of the percent correct to a linear scale with a mean of 13 and a standard deviation of 4 and is used as an index of item difficulty by ETS test development staff. The $\hat{\Delta}$ statistic is interpreted as a difference in the difficulty of items for the reference and focal groups on the delta scale (Zieky, 1993).

The Mantel-Haenszel DIF Parameter

The goal of this section is to derive an expression for a theoretical MH DIF parameter, Δ, that represents the expected value for $\hat{\Delta}$ that would be obtained for a studied item if examinees could be matched on θ exactly. The examinees' θ's will be assumed to be sampled from an infinitely large population with continuous θ density functions for the reference and focal group populations and with a fixed ratio of reference group population size to focal group population size. Such assumptions are typical of DIF simulation studies.

The general approach taken in this derivation is to assume that examinees are matched on a carefully chosen theoretical matching test, then to carefully let test length go to infinity to result in the desired matching of examinees on θ. The derivation begins by assuming examinees are matched based on their scores on a theoretical matching test (rather than a real or simulated one) that sorts examinees by test scores exactly (i.e., with no error). Assuming a matching test that has perfect reliability allows the convenience of not including the score on the studied item in the matching criterion. The derivation could just as well be carried out with the score on the studied item included in the matching criterion and would, in the end, result in the same formula for the MH DIF parameter. However, the derivation is less cumbersome when the score on the studied item is not included in the matching criterion.

Consider a matching test consisting of S Guttman items (Guttman, 1944) that span the ability range from $-\sqrt{S}$ to $\sqrt{S}$ at equal intervals for both the reference and focal groups. Let the difficulty parameters of these items be denoted by β_i and assume that they are on the same scale as the θ's. The ordered difficulty parameters will be denoted by $\beta_{(i)}$, where $\beta_{(i+1)} - \beta_{(i)} = \delta$ for all $i = 1$ to S. Because the items are equally spaced over the

[1] $\hat{\alpha}_s$ will be difficult to estimate when either I_{Rs} is close to 0 or C_{Fs} is close to 0. In such cases, $\hat{\alpha}_s$ will be very unstable and cause large variability in the estimation of $\hat{\alpha}$. Thus when estimating a constant odds ratio, it makes sense to use an estimator that gives proportionally less weight to score cells according to how close the cell is to having $I_{Rs} = 0$ or $C_{Fs} = 0$. By using weights proportional to $I_{Rs}C_{Fs}$, the Mantel-Haenszel odds ratio accomplishes this objective. Thus the weights in Equation 5 will be small for cells where $\hat{\alpha}_s$ will be unstable.

above stated ability range, we obtain $\delta = 2\sqrt{S}/S = 2/\sqrt{S}$. If an examinee obtains a score of s, this indicates that the examinee's θ was at least $\beta_{(s)}$, but less than $\beta_{(s)} + \delta$. Thus, the probability of an examinee with ability θ obtaining a score of exactly s, denoted by $P(s|\theta)$, is then given by

$$P(s|\theta) = 1 \text{ if } \beta_{(s)} \leq \theta < \beta_{(s)} + \delta, \text{ and} \tag{7}$$

$$P(s|\theta) = 0 \text{ otherwise.}$$

Now consider a specific cell C_{Rs} in the 2 x 2 x S contingency table (see Table 1). The theoretical probability of a particular examinee from the reference group, with ability θ, contributing to this cell is given by

$$P(s \mid \theta) P_R(\theta), \tag{8}$$

where $P_R(\theta)$ is the reference group IRF for the studied item, as defined previously, which is not assumed to take any particular functional form for either the reference group or focal group. By assuming that the reference group θ's follow some underlying distribution $f_R(\theta)$, we can then determine the expected total cell count for C_{Rs}, for a sample of N_R reference group examinees from the following integral:

$$E[C_{Rs}] = N_R \int_{-\infty}^{\infty} P(s \mid \theta) P_R(\theta) f_R(\theta) d\theta . \tag{9}$$

Similar equations can be developed for all the other cells in the 2 x 2 x S contingency table. Because the only θ's that contribute to C_{Rs} are between $\beta_{(s)}$ and $\beta_{(s)} + \delta$, and because $P(s|\theta)$ is unity in this range of θ's, Equation 9 can be written as

$$E[C_{Rs}] = N_R \int_{\beta_{(s)}}^{\beta_{(s)}+\delta} P_R(\theta) f_R(\theta) d\theta . \tag{10}$$

This expression may be simplified by invoking the mean-value theorem to replace the integral with a product. According to the mean-value theorem, there exists some value $\theta'_{C_{Rs}}$ in the interval $\beta_{(s)} \leq \theta'_{C_{Rs}} < \beta_{(s)} + \delta$ that, when evaluated in the integrand expression [$P_R(\theta) f_R(\theta)$, in this case], then multiplied by the width of the interval (δ, in this case), will result in the same value as the original integral. Thus, by applying the mean-value theorem to Equation 10, E[C_{Rs}] can be written as

$$E[C_{Rs}] = N_R P_R(\theta'_{C_{Rs}}) f_R(\theta'_{C_{Rs}}) \delta , \tag{11}$$

for some $\beta_{(s)} \leq \theta'_{C_{Rs}} < \beta_{(s)} + \delta$.

Similar expressions can be derived for E[C_{Fs}], E[I_{Rs}], and E[I_{Fs}], with corresponding ability mean values $\theta'_{C_{Fs}}$, $\theta'_{I_{Rs}}$, and $\theta'_{I_{Fs}}$. The expression for E[$N_{Total,s}$], which is required for the MH weights, is given by

$$E[N_{Total,s}] = N_{Total}\{\gamma_F[P_F(\theta'_{C_{Fs}})f_F(\theta'_{C_{Fs}}) + Q_F(\theta'_{I_{Fs}})f_F(\theta'_{I_{Fs}})] + \gamma_R[P_R(\theta'_{C_{Rs}})f_R(\theta'_{C_{Rs}}) + Q_R(\theta'_{I_{Rs}})f_R(\theta'_{I_{Rs}})]\}\delta \quad (12)$$

where $N_{Total} = N_R + N_F$, the total number of examinees,

γ_F is the proportion of the total number of examinees who are in the focal group, and
γ_R is the proportion in the reference group.

An equation for α, which is analogous to the equation of $\hat{\alpha}$ (Equation 5), can then be specified by substituting in these expected values and simplifying, giving

$$\alpha = \frac{\sum_s \left(\frac{Q_R(\theta'_{I_{Rs}})f_R(\theta'_{I_{Rs}})P_F(\theta'_{C_{Fs}})f_F(\theta'_{C_{Fs}})\delta}{\gamma_F[P_F(\theta'_{C_{Fs}})f_F(\theta'_{C_{Fs}}) + Q_F(\theta'_{I_{Fs}})f_F(\theta'_{I_{Fs}})] + \gamma_R[P_R(\theta'_{C_{Rs}})f_R(\theta'_{C_{Rs}}) + Q_R(\theta'_{I_{Rs}})f_R(\theta'_{I_{Rs}})]} \right) \alpha(s)}{\sum_s \left(\frac{Q_R(\theta'_{I_{Rs}})f_R(\theta'_{I_{Rs}})P_F(\theta'_{C_{Fs}})f_F(\theta'_{C_{Fs}})\delta}{\gamma_F[P_F(\theta'_{C_{Fs}})f_F(\theta'_{C_{Fs}}) + Q_F(\theta'_{I_{Fs}})f_F(\theta'_{I_{Fs}})] + \gamma_R[P_R(\theta'_{C_{Rs}})f_R(\theta'_{C_{Rs}}) + Q_R(\theta'_{I_{Rs}})f_R(\theta'_{I_{Rs}})]} \right)} \quad (13)$$

where

$$\alpha(s) = \frac{P_R(\theta'_{C_{R_s}})Q_F(\theta'_{I_{F_s}})}{P_F(\theta'_{C_{F_s}})Q_R(\theta'_{I_{R_s}})}. \quad (14)$$

Finally, we let test length, S, become asymptotically large while maintaining the equally spaced Guttman items and the difficulty parameter range of $-\sqrt{S}$ to $\sqrt{S}$. Hence, the θ range approaches $-\infty$ to $+\infty$, the width of the intervals (δ) approaches 0, and $\theta'_{C_{Rs}}$, $\theta'_{C_{Fs}}$, $\theta'_{I_{Rs}}$, and $\theta'_{I_{Fs}}$ all approach the same value, say θ_S. Because score s approaches a continuous variable, θ_S may be replaced with θ, and the summations over the δ intervals in Equation 13 can then be replaced by integrals over $d\theta$. Because all the different θ' values for the same score s approach a common value, the denominator of the MH weights simplifies to $\gamma_F f_F(\theta) + \gamma_R f_R(\theta)$. Thus, our final formula for α [which applies weights to $\alpha(\theta)$ in an analogous manner as the MH odds ratio applies weights to $\hat{\alpha}_s$] is given by

$$\alpha = \frac{\int_{-\infty}^{\infty} Q_R(\theta)P_F(\theta)\frac{f_R(\theta)f_F(\theta)}{\gamma_F f_F(\theta) + \gamma_R f_R(\theta)}\alpha(\theta)d\theta}{\int_{-\infty}^{\infty} Q_R(\theta)P_F(\theta)\frac{f_R(\theta)f_F(\theta)}{\gamma_F f_F(\theta) + \gamma_R f_R(\theta)}d\theta}. \quad (15)$$

This formula is the same as the one postulated by Spray and Miller (1992).

Note that $\alpha(\theta)$ reduces to Equation 2 in the case of 1PL and 2PL, and it reduces to Equation 3 in the case of 3PL. This quantity is substituted into the final formula for the MH DIF parameter,

$$\Delta = -2.35\ ln(\alpha). \tag{16}$$

Although no closed-form solution exists for the general solution of the integral for α (Equation 15), numerical integration can be used to compute α. We developed software for the computation of Δ, via α using Equation 15, by employing a standard technique, Simpson's 1/3 Rule [see, for example, Gerald, 1970] and using 199 integration points across a θ range from –6 to 6. This software for calculating the MH DIF population parameter is available upon request from the authors.

Recall that when the IRF of the studied item follows either the 2PL or 1PL model and DIF is uniform, $\alpha(\theta)$ is a constant with respect to θ, as given in Equation 2, and comes outside the integral in the numerator in Equation 15. The integrals in the numerator and denominator then cancel each other out, which results in

$$\alpha = e^{-1.7a(b_R - b_F)} \tag{17}$$

and, consequently,

$$\Delta = -2.35ln(e^{-1.7a(b_R - b_F)}) = -2.35(-1.7a(b_R - b_F)) = 4a(b_R - b_F), \tag{18}$$

which is the well-known form for Δ in the 1PL and 2PL cases (e.g., Donoghue, Holland, & Thayer, 1993). The estimate of Δ is interpreted as an estimate of the difference in difficulty level for the reference and focal groups on the studied item as measured on the ETS delta scale (Zieky, 1993). In the case of 1PL, this interpretation of $\hat{\Delta}$ is justified because Δ is $4(b_R - b_F)$, a difference in b's multiplied by the appropriate constant. The interpretation of $\hat{\Delta}$ in the 2PL case is less direct because of the a parameter in the $4a(b_R - b_F)$ formula.

In the 3PL case, the odds ratio parameter $\alpha(\theta)$ is not constant across θ [see Equation 3], thus the simple $4a(b_R - b_F)$ rule does not apply. In this case, because $\alpha(\theta)$ is not constant across θ; and because Δ does not have a simple form, it is not clear what the interpretation of $\hat{\Delta}$ should be.

Because $\hat{\Delta}$ has become the industry standard (and the most widely used) DIF estimator, and it is used with data that is modeled well by the 3PL IRF, it is important to know whether the interpretation of $\hat{\Delta}$ as an estimator of difference in difficulty level holds approximately true when the studied item is modeled with 3PL uniform DIF.

Computation of Δ Under Realistic Conditions

Our formulation of the MH DIF parameter was calculated for numerous uniform DIF conditions. The θ's for the reference group were always specified as following a standard normal distribution. The θ's for the focal group were specified as following a normal distribution with unit variance; the mean was set at –1, –.5, 0, .5, or 1 (five levels). The ratio of the reference group size to the focal group size (Ratio) was 5, 4, 3, 2, or 1 (five levels). The discrimination parameter (a) was set at .5, .7, .9, 1.1, or 1.3 (five levels). The lower asymptote (c) was set at 0 (the 2PL case), .1, .2, or .3 (four levels). The difficulty parameter for the reference group (b_R) was set at –2, –1.5, –1, –.5, 0, .5, 1, 1.5, or 2 (nine levels). The difficulty parameter for the focal group (b_F) was set equal to b_R (i.e., no DIF) and to $b_R \pm .1, .2, ..., 1.5$ (31 levels). (That is, the difference in b values for the two groups was $0, \pm .1, \pm .2, ..., \pm 1.5$.) The item parameter factors were fully crossed, resulting in 5 x 4 x 9 x 31 = 5,580 "items" for which Δ was calculated (levels of a by c by b_R by b_F) for the 5 x 5 = 25 examinee combinations (focal group mean by ratio of reference to focal group size). A representative subset of the results will be presented.

ETS has developed a classification scheme to flag items with DIF (see Zieky, 1993), which is also used at LSAC. Items where $|\hat{\Delta}|$ is at least 1.5 and significantly greater than 1.0 are classified as "C" items, or moderate to large DIF. Items with a flag of C are routinely excluded from test assembly at LSAC. Items where $|\hat{\Delta}|$ is not significantly greater than 0.0 or $|\hat{\Delta}|$ is less than 1.0 are classified as "A" items, or negligible DIF. All other items are classified as "B" items, or slight to moderate DIF. Note that in the 1PL case, according to the equation $\Delta = 4(b_R - b_F)$, a difference in b values of .375 or more would result in $|\Delta|$ values corresponding to a C flag ($\Delta \geq 1.5$). Thus 24 of the 31 differences in b values used in the present study represent substantial DIF levels ($b_R - b_F = \pm .4, \pm .5, ..., \pm 1.5$). (Statistical significance is not relevant here because we are calculating the parameter itself, not a statistic.)

The panels in Figure 2 show how c affects Δ at the various values of b_R and b_F. For all items Ratio = 5, $a = 0.9$ and focal group mean = –1.0. The values of b_F are plotted along the horizontal axis, and Δ is plotted along the vertical axis. The values of b_R are indicated by the different symbols. For instance, all solid circles represent $b_R = 2.0$, and the position of a particular solid circle in reference to the horizontal axis indicates the value of b_F. When there is no difference in difficulty for the two groups ($b_R = b_F$) for all levels of b_R, $\Delta = 0$ (no DIF), as expected. The horizontal lines in the figure at Δ values of $\pm$ 1.5 and $\pm$ 1.0 delineate the Δ cutoff values for C and B DIF items, respectively.

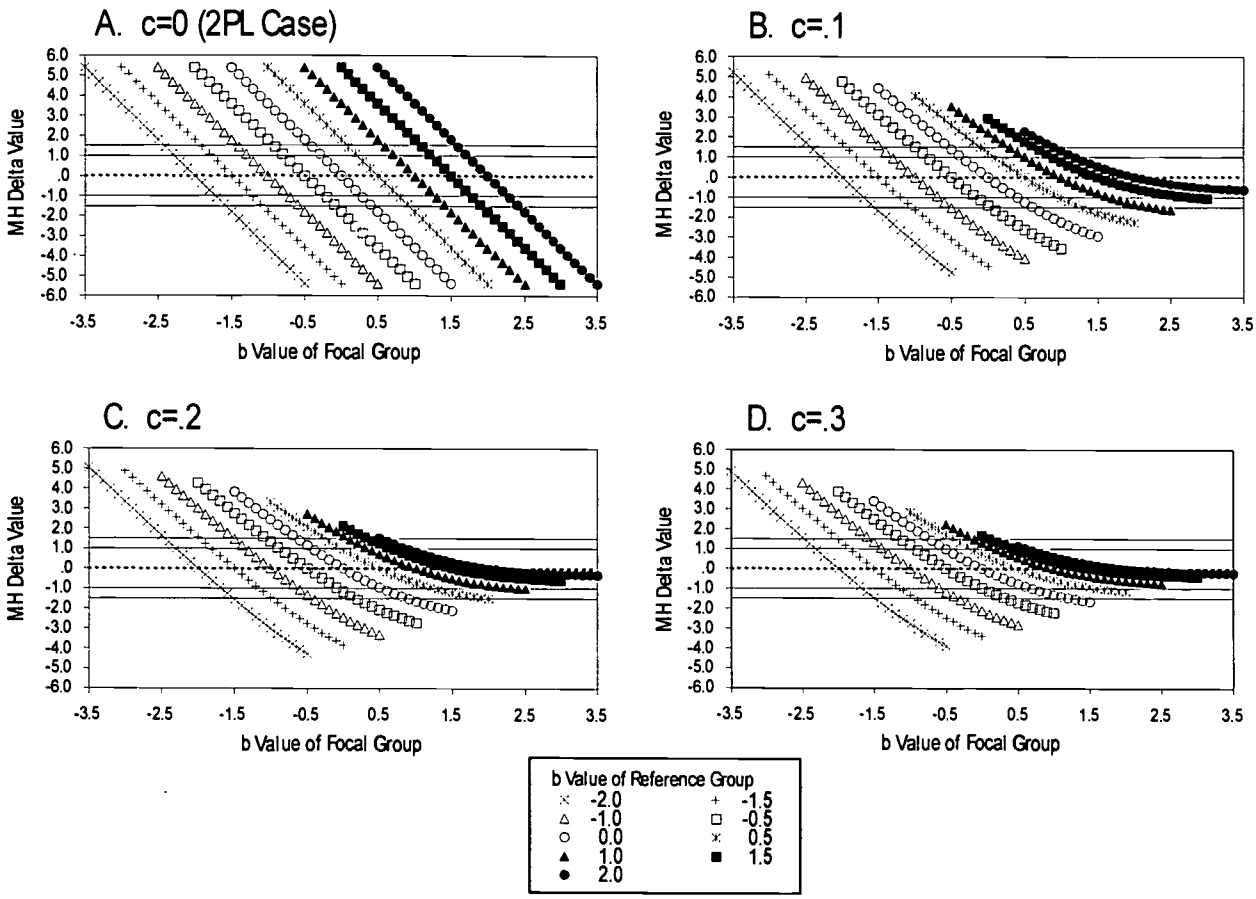

FIGURE 2. *Variation of Δ with b_R (b value of reference group, indicated by the symbol), b_F (b value of focal group, indicated along the horizontal axis), and c (which varies across panels) for fixed values of a = .9. Ratio = 5, and focal group mean = –1.0.*

BEST COPY AVAILABLE

For $c = 0$ (the 2PL case, shown in panel A of Figure 2), we find, as expected, that $\Delta = 4a(b_R - b_F)$. In the 1PL case ($a = 1$ and $c = 0$), $b_R - b_F = 0.4$ would result in a C level value of Δ. However, for the items in panel A of Figure 2, ($a = .9$), it is clear that when $a \neq 1$, Δ is not merely a transformation of $b_R - b_F$ on to the ETS delta scale as it is usually interpreted. This effect of a on the interpretation of Δ in the 2PL case can be seen more clearly if we compare what Δ would be for different values of a for the same value of $b_R - b_F$. For example, if a were set equal to 1.3, a value of $b_R - b_F$ of about 0.29 yields a Δ of 1.5, whereas if a were set equal to 0.5, the same $b_R - b_F$ of 0.29 results in a Δ of only 0.58 in the 2PL case. Even though $\Delta = 1.5$ (when $a = 1.3$) is interpreted as an indication of a larger difference in difficulty than $\Delta = 0.58$ (when $a = 0.5$), the difficulty differences are in actuality exactly the same ($b_R - b_F = 0.29$). As interesting as these 2PL results are, they are, for the most part, merely didactic because item responses on the Law School Admission Test (LSAT) and other nationally administered standardized tests often follow the 3PL model due to the extensive use of multiple-choice items on these tests. Therefore, we now turn our attention to the results for the 3PL model.

Panels B, C, and D of Figure 2 show that the introduction of positive values of c into the IRF model has quite a large effect on Δ. Clearly, the 2PL formula, $\Delta = 4a(b_R - b_F)$, is not a very good general approximation to Δ in the case of 3PL uniform DIF. Except for the very lowest values of b_R, Δ is moderately to substantially reduced from its 2PL value as b_R increases, especially when the item favors the reference group (i.e., negative $b_R - b_F$ values, and hence Δ values, indicating the item is more difficult for the focal group). Even for b_R values as small as 0, which is usually close to the mean of all the b values, the 3PL value of Δ is on average less than half the corresponding 2PL value when averaged over the 15 $b_R - b_F$ values for DIF against the focal group for the case of $c = 0.2$ (approximately the average c value on the LSAT). Even when such items display substantial amounts of DIF against the focal group ($b_R - b_F = -1.0$ to -1.5), Δ seldom reaches the level of a "B" item value, much less a "C" item.

The panels in Figure 3 show how varying a affects Δ at the various values of b_R and b_F in the 3PL case ($c = 0.2$). For all items Ratio = 5 and focal group mean = –1.0. The primary pattern of Δ decreasing when b_R increases, which was evident in Figure 2 is maintained across varying values of a, as shown in Figure 3. Additionally, varying a does have a large effect on Δ, as expected from the 2PL results discussed above. As was seen in Figure 2, the 3PL Δ values for the lowest values of b_R followed the 2PL values [$\Delta = 4a(b_R - b_F)$] more closely than for the higher values of b_R for all levels of a. However, as a increases, the distortion in Δ becomes more dramatic as b_R increases. Thus, the effect of varying a on Δ for low values of b_R, was similar to that in the 2PL case—as a increases so does Δ. However, for higher values of b_R, the effect of varying a was quite unexpectedly the *opposite* of what would be predicted by the 2PL formula—as a increased, Δ *decreased*, with the effect occurring most strongly for the case of DIF against the focal group. For example, when $b_R = 1$ and $b_F = 2$, at $a = 0.5$, 0.7, 0.9, 1.1, and 1.3, the respective Δ values are –0.957, –0.783, –0.603, –0.466, and –0.369.

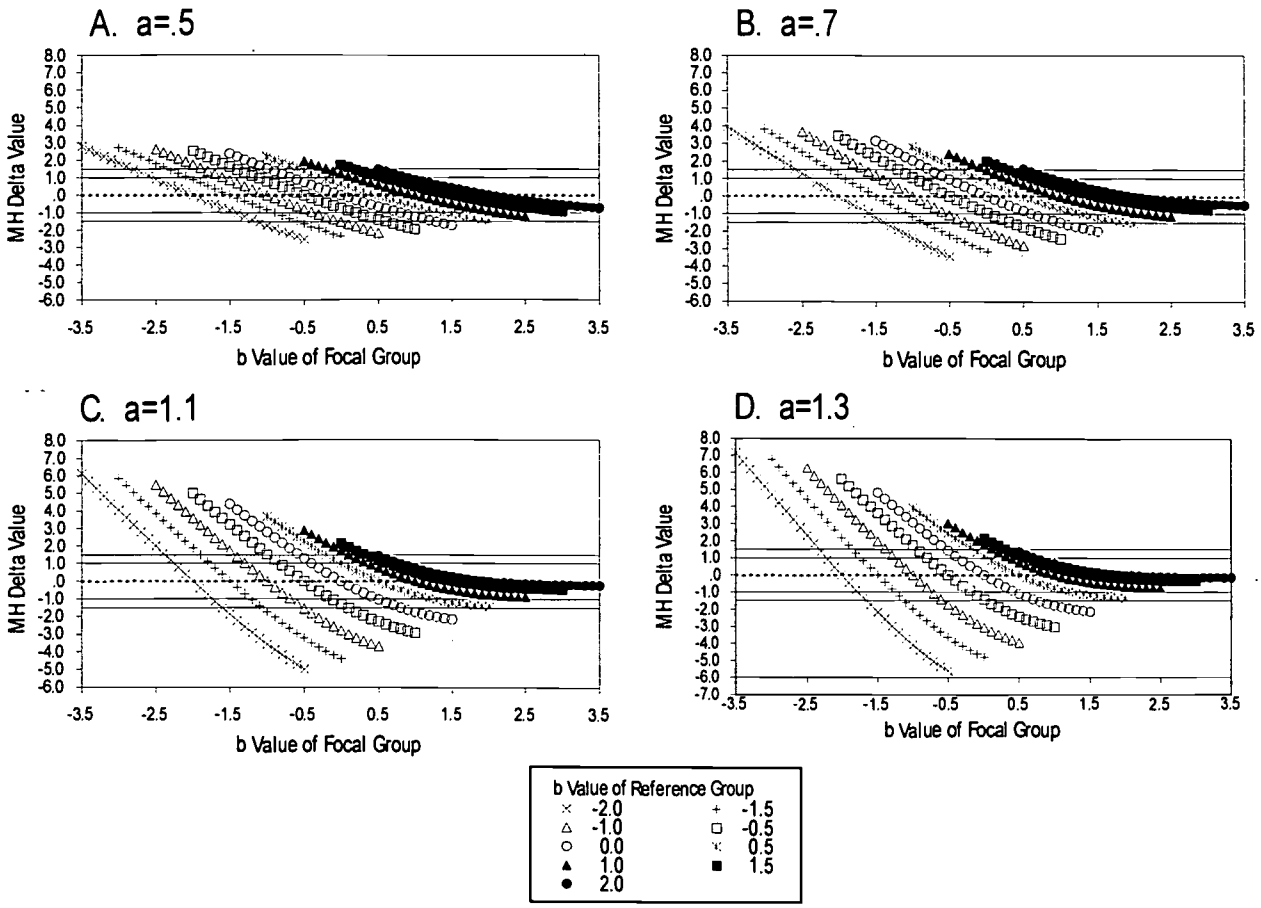

FIGURE 3. *Variation of Δ with b_R (b value of reference group, indicated by the symbol), b_F (b value of focal group, indicated along the horizontal axis), and a (which varies across panels) for fixed values of c = .2, Ratio = 5, and focal group mean = –1.0.*

The panels in Figure 4 show how varying the focal group mean affects Δ at the various values of b_R and b_F in the 3PL case ($c = 0.2$) for Ratio = 5 and $a = 0.9$. In all cases, the focal group θ's were specified as following a normal distribution with unit standard deviation; the mean of the distribution used the values –1 (shown in Panel C of Figure 2), –.5, 0, .5, and 1 (shown in Panels A through D of Figure 4). Recall that the reference group's θ's were always specified as following a standard normal distribution. As in Figures 2 and 3, the most dramatic effect is that Δ decreases as b_R increases. Figure 4 shows that varying the focal group mean has a noticeable effect on Δ, a result that the 2PL formula for Δ could not predict. The results show that the focal groups for which Δ is most reduced relative to its 2PL value are the groups with the lowest proficiency distributions (focal means of –1.0, –0.5, and 0.0; the focal group mean of –1.0 is shown in Panel C of Figure 2). Even though the results are better (i.e., Δ is distorted less compared to the 2PL formula) for the focal groups with means greater than 0, the shrinkage effect remains quite strong when $b_R \geq 1.0$ (for the focal mean of 0.5) or when $b_R \geq 1.5$ (for the focal mean of 1.0). Moreover, the results for Figure 4 are only for $a = 0.9$, and Figure 3 showed that higher a values cause even larger distortions in Δ. The focal groups used in DIF analyses at LSAC typically have estimated means that are lower than the reference group's mean; thus we clearly deal primarily with the larger distortions in Δ that result when the focal group's mean is less than the reference group's mean.

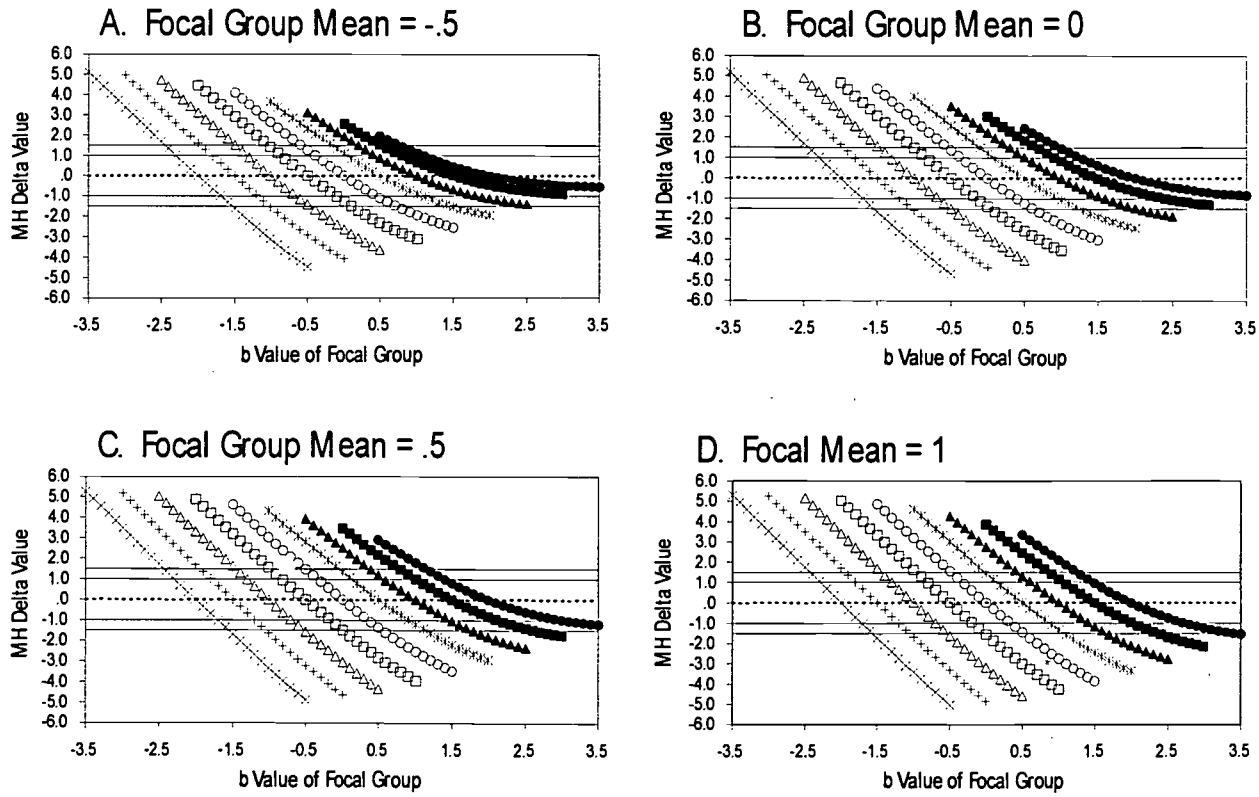

FIGURE 4. *Variation of* Δ *with* b_R *(b value of reference group, indicated by the symbol),* b_F *(b value of focal group, indicated along the horizontal axis), and the mean of the focal group (which varies across panels) for fixed values of* $c = .2$, $a = .9$, *and Ratio = 5.*

The panels in Figure 5 show how varying the ratio of reference- to focal-group size affects Δ at the various values of b_R and b_F in the 3PL case ($c = 0.2$) for focal group mean = –1 and $a = 0.9$. Again, the most dramatic effect is that Δ decreases as b_R increases. As indicated in Figure 5, varying the ratio of reference group size to focal group size has a relatively minor effect on Δ.

BEST COPY AVAILABLE

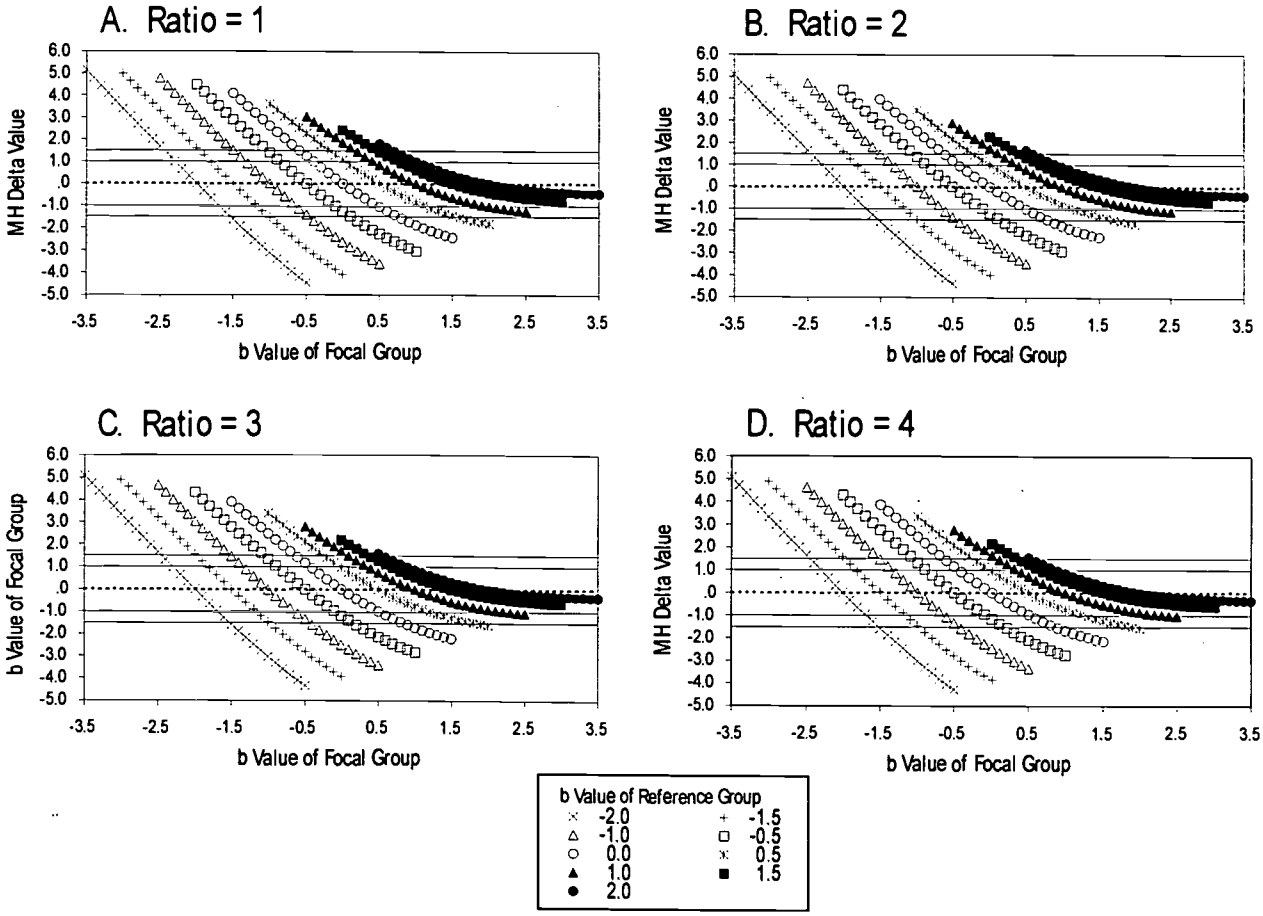

FIGURE 5. *Variation of* Δ *with* b_R *(b value of reference group, indicated by the symbol),* b_F *(b value of focal group, indicated along the horizontal axis), and the ratio of the reference group size to the focal group size (which varies across panels) for fixed values of* $c = .2$, $a = .9$, *and focal group mean* $= -1.0$.

Correspondence to Simulated Data Results

Previous simulation studies by Allen and Donoghue (1996); Donoghue, Holland, and Thayer (1993); and Uttaro and Millsap (1994) have shown an unexplained tendency of the MH DIF estimator to decrease with difficulty level. The above results indicate that this tendency may be explained, at least in part, by the behavior of the MH DIF parameter itself. As an example, Table 2 shows the mean $\hat{\Delta}$ values (in bold) from Allen and Donoghue (1996, their Table 4) alongside our MH DIF parameter values (the Δ's, also in bold) where α was calculated from Equation 15. Various values of a and b_R are shown. For all items, $c = .2$, $b_F = b_R + .4$. As can be seen in Table 2, our MH DIF parameter values, Δ, reproduce Allen and Donoghue's DIF estimator means (labelled "Mean $\hat{\Delta}$" in the table) quite well. The amount that the parameter values differ from the estimated values can probably be attributed to the known estimation bias of the MH DIF estimator that occurs when the reference and focal group populations have a large difference in their proficiency means (as was the case for the data simulated in Allen and Donoghue, 1996; this bias is evident in the "No DIF" column—all of the $\hat{\Delta}$ means would be within a standard error[2] or two of 0 if there was no bias).

[2] Standard error is estimated by dividing the tabulated standard deviation (SD) by the square root of 150, the number of replications.

TABLE 2

Comparison of MH DIF estimates from Allen and Donoghue (1996) with our MH DIF parameter, Δ

		No DIF ($b_F = b_R$)		DIF ($b_F = b_R + .4$)		
a	b_R	Mean $\hat{\Delta}$	SD of $\hat{\Delta}$	Mean $\hat{\Delta}$	SD of $\hat{\Delta}$	Δ
.5	-2	-.02	.22	**-.73**	.23	**-.73**
.5	-1	.00	.20	**-.66**	.19	**-.67**
.5	0	.04	.18	**-.52**	.19	**-.58**
.5	1	.06	.17	**-.36**	.18	**-.44**
.5	2	.08	.20	**-.18**	.20	**-.29**
1.0	-2	-.28	.33	**-1.71**	.28	**-1.49**
1.0	-1	-.17	.24	**-1.42**	.20	**-1.30**
1.0	0	-.02	.19	**-.96**	.18	**-.95**
1.0	1	.07	.19	**-.43**	.22	**-.48**
1.0	2	.14	.19	**-.00**	.22	**-.15**
1.5	-2	-.54	.46	**-2.63**	.38	**-2.21**
1.5	-1	-.28	.23	**-2.07**	.22	**-1.85**
1.5	0	-.04	.21	**-1.17**	.21	**-1.17**
1.5	1	.09	.21	**-.31**	.22	**-.40**
1.5	2	.14	.24	**.05**	.23	**-.06**

Note. In the calculation of mean and standard deviation of $\hat{\Delta}$, Allen and Donoghue (1996) used 150 replications. The "No DIF" estimates from Allen and Donoghue indicate the amount of bias present in $\hat{\Delta}$ caused by the large difference in mean proficiency between the reference and focal groups. In all cases, $c = .2$. The reference group's θ's were sampled from a normal distribution with a mean of 0 and standard deviation of .7. The focal group's θ's were sampled from a normal distribution with a mean of $-.7$ and standard deviation of .8. The ratio of reference group size (5,100) to focal group size (1,050) was 4.857. These same θ distributions and ratio were specified for the calculation of Δ, given by $\Delta = -2.35ln(\alpha)$, where α is given in Equation 15.

Discussion

Before items are ever presented to examinees on an LSAT form, the items undergo an extensive sensitivity review process. Despite these precautions, some items may function differently in various subgroups (i.e., exhibit DIF). DIF statistics are designed to detect such items. Several DIF statistics have been developed, but the MH DIF procedure, which is used at LSAC, has become the most widely used methodology and is recognized as the testing industry standard. Although the behavior of the MH DIF estimator's population parameter is known in 1PL and 2PL data, it has not been known in 3PL data because the formulation of the MH population parameter, Δ, has been an unsolved problem in the 3PL case. This lack of knowledge about Δ for 3PL items has limited the evaluation of the statistical bias in $\hat{\Delta}$ and has also hindered the understanding of the observed effects of simulation study factors on $\hat{\Delta}$. In particular, several researchers have found that the difficulty level of a 3PL DIF item can have a sizable effect on the magnitude of $\hat{\Delta}$ (Allen & Donoghue, 1996; Donoghue, Holland, & Thayer, 1993; Uttaro & Millsap, 1994), but none of these studies could adequately explain the cause of this effect.

The present statistical report formulated a population DIF parameter for the MH DIF estimator for any IRF model, including the 3PL model, and investigated its behavior with respect to a number of examinee and item factors through a systematic set of calculations. The findings presented here indicate that caution should be used in applying the MH DIF estimator to item response data that follow the 3PL model. In particular, the results indicate that the MH DIF estimator may exhibit reduced statistical power to detect DIF in 3PL items of medium or high difficulty, even when DIF is substantial (i.e., large difference in b's), especially when the focal group has a low mean proficiency. Additionally, it was shown that the behavior of the MH DIF population parameter can account for the unexplained behavior of $\hat{\Delta}$ with respect to difficulty level observed in a past simulation study (Allen & Donoghue, 1996). The fact that $\hat{\Delta}$ is smaller (in absolute value) than expected for 3PL items of

BEST COPY AVAILABLE

medium or high difficulty, as compared with 1PL or 2PL items, now can be explained because the value of the MH DIF parameter, Δ, also exhibits this pattern. Thus $\hat{\Delta}$ should be used with caution until the apparent deficiencies of this procedure are better understood or corrected.

The implications of this study on the routine operational task of identifying DIF at LSAC are still unknown because real data do not mimic simulated data exactly. However, because some items on the LSAT are known to exhibit guessing behavior, the results certainly suggest that additional research is warranted.

References

Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33*, 231-251.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Menlo Park, CA: Addison-Wesley.

Donoghue, J., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum.

Gerald, C. F. (1970). *Applied numerical analysis*. Reading, MA: Addison-Wesley.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*, 215-230.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Spray, J.A., & Miller, T.R. (1992). *Performance of the Mantel-Haenszel statistic and the standardized difference in proportion correct when population ability distributions are incongruent* (Research Report No. 92-01). Iowa City, IA: American College Testing.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*, 15-25.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").